Unlocking Efficiency in Large Language Model Inference: A Comprehensive Survey of Speculative Decoding

Year: 2024 | *Citations: 194* | *Authors: Heming Xia, Zhe Yang, Qingxiu Dong*

Abstract

To mitigate the high inference latency stemming from autoregressive decoding in Large Language Models (LLMs), Speculative Decoding has emerged as a novel decoding paradigm for LLM inference. In each decoding step, this method first drafts several future tokens efficiently and then verifies them in parallel. Unlike autoregressive decoding, Speculative Decoding facilitates the simultaneous decoding of multiple tokens per step, thereby accelerating inference. This paper presents a comprehensive overview and analysis of this promising decoding paradigm. We begin by providing a formal definition and formulation of Speculative Decoding. Then, we organize in-depth discussions on its key facets, such as drafter selection and verification strategies. Furthermore, we present a comparative analysis of leading methods under third-party testing environments. We aim for this work to serve as a catalyst for further research on Speculative Decoding, ultimately contributing to more efficient LLM inference.